Journal of Engineering and Applied Sciences 14 (5): 1662-1667, 2019
ISSN: 1816-949X
© Medwell Journals, 2019

An Economic and Mathematical Model of IT Service Provider Selection on the Basis of Analysis of Non-Structured Text Documents

[1]Maxim Dli, [2]Nikolai Salov, [1]Tatyana Kakatunova and [1]Dmitrii Tukaev
[1]Department of Management and Information Technologies in Economics,
Smolensk Branch, National Research University "MPEI", Smolensk, Russian Federation
[2]Department of Economics in Power Engineering and Industry,
National Research University "MPEI", Moscow, Russian Federation

Abstract: The study proposes an economic and mathematical IT service provider selection model requiring no preliminary intellectual processing of inputs in contrast to known models. In addition, this approach is characterized by lower computational complexity in contrast to known topic modeling-based recommendation algorithms and enables taking into account the similarity of various topics.

Key words: IT outsourcing, analysis of non-structured text documents, text document clustering, recommendation systems, topic modeling, additive regularization of topic models

INTRODUCTION

The selection of an IT provider for works associated with a software development and maintenance project involves the processing of large volumes of noisy data. It is due to globalization and available information volume growth caused by information technology development. The prerequisite for making an organization competitive in such conditions is using filtering methods, i.e., highlighting the providers alone that are relevant for the job among their totality. Such filtering tasks are expedient to be solved using recommendation systems.

A characteristic feature of this task is that the information on works and providers is semistructured, that is represented in the form of textual descriptions (organization descriptions, work specifications). Given the above said, the goal of this research seems to be of interest today as it is to develop an economic and mathematical model of IT service provider selection on the basis of analysis of non-structured text documents. The following tasks are to be performed to achieve the goal set:

- To review known recommendation generation methods used in this field
- To adapt the topic model for ranking elements by degree of target entity relevance
- To assess the ranking quality

Literature review: Approaches based on in-memory approaches (collaborative content filtration methods as well as their hybrid modifications) are prevalent at present in building recommendation systems. Examples of such systems are described in many works (Pazzani and Billsus, 2007; Burke, 1999; Ghazanfar and Prugel-Bennett, 2010). Despite the advantage of such approaches mainly due to simple implementation and interpretation, they have a series of significant disadvantages such as 'cold start' problem, sparsity of large data volumes and lack of theoretical substantiation of approach combination in hybrid versions (Gomzin and Korshunov, 2012).

Approaches based on attribute extraction from various presentations of IT service and research providers are another widely used group of approaches to building recommendation systems (Al-Otaibi and Ykhlef, 2012). These approaches have no in-memory system disadvantages and often generate recommendations of better quality. Nevertheless, the use of these methods in industrial applications presents certain difficulties because of attribute generation and selection process complexity as well as high requirements for the learning sample.

It should be taken into consideration at the same time that some recommendation generation algorithms are commercial secrets and their building technology us unknown to a wide circle of persons.

MATERIALS AND METHODS

The combined cluster structure of IT service and work providers under the system software development and maintenance project can be restored on the basis of

Corresponding Author: Nikolai Salov, Department of Economics in Power Engineering and Industry,
National Research University "MPEI", Moscow, Russian Federation

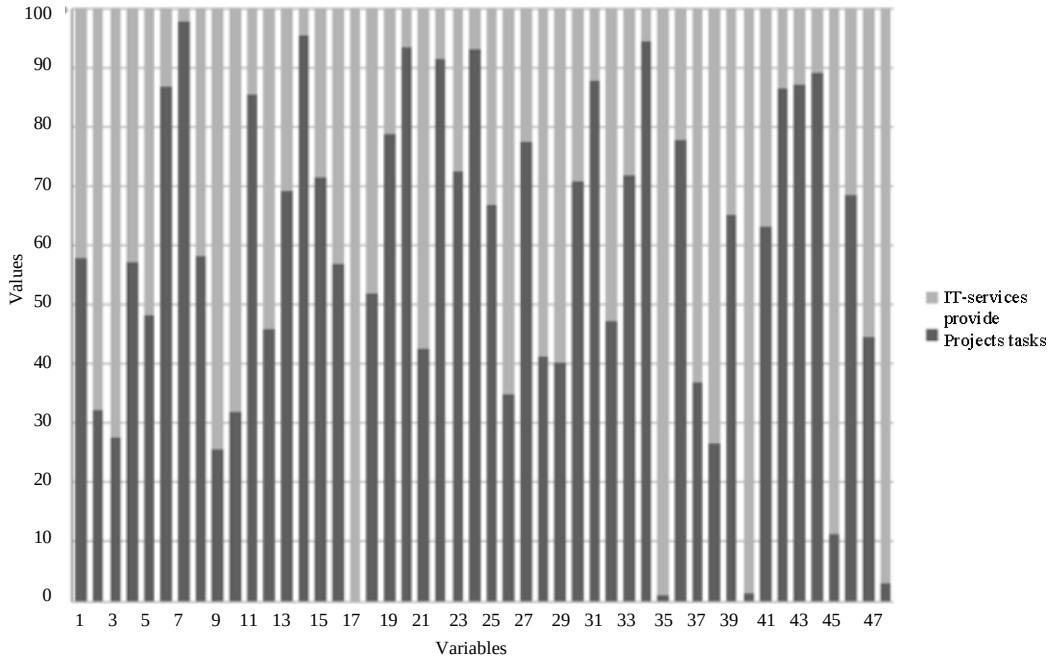

Fig. 1: Example of distribution of textual project IT service and work provider descriptions by cluster

their textual descriptions (Salov, 2017). Figure 1 shows an example of distribution of appropriate documents by cluster.

In this connection, the researchers are proposed to use topic modeling (Vorontsov and Potapenko, 2012) for the development of IT service provider search recommendation algorithm. The recommendation generation diagram is presented in Fig. 2.

Probabilistic topic models enable handling of documents described in natural languages. It enables using such models without necessity of preliminary intellectual data processing. Nevertheless, the inputs should be processed in a certain way (tokenization, stop word filtration, lemmatization, n-Gram highlighting) but all such operations are standardized and are performed without expert's direct participation. An analysis of open data from most popular recruiting services shows that textual descriptions of actors and tasks (d^1, ..., d^n) are divided in several semantically different components, i.e., title, description и skills. It seems expedient to divide the components (length, background word content) at model input to take into consideration their specific features which was implemented by highlighting three modalities in the topic model. The special features of components are taken into consideration by specifying a separate term matrix Φ_m for each modality in this case (Ianina *et al*., 2017).

Solution nonuniqueness and instability is the disadvantage of probabilistic topic models. Regularization is one of solutions to this problem. An additive Regularization based Approach to Topic Model building (ARTM) is of the greatest interest from this point of view. The idea behind this approach is to impose additional restrictions on the solution sought, each being formalized in the form of optimization criterion regularizer $R_i(\Phi, \theta)\rightarrow max$ depending on model parameters. The weighted sum of such criteria $R(\Phi, \theta)$ is maximized jointly with the basic likelihood criterion (Vorontsov, 2014). The advantages of this approach include simple mathematical apparatus, generalization of the majority of known approaches and software implementation availability (Kochedykov *et al*., 2017). The disadvantage of the ARTM approach is the necessity of manually selecting the regularization route in each specific case because of necessity of searching and subsequently analyzing a large number of various regularizer value combinations. Given the above said, the researcher suggests to use the additive regularization of topic models for building IT service provider search recommendation algorithm and the maximization of weighted normalized sum of internal topic model quality criteria for regularization strategy selection in this study (Salov, 2017).

Topic model building results in matrix $\theta(\theta(d_1), ..., \theta(d_n))$ describing topics by documents (textual descriptions of project works and IT service providers) 1, ..., n as well as function $\varphi(d^*)\rightarrow\theta(d^*)$ providing the topic distribution in new document d* (that did not participate in model building). Given that, the task of generating

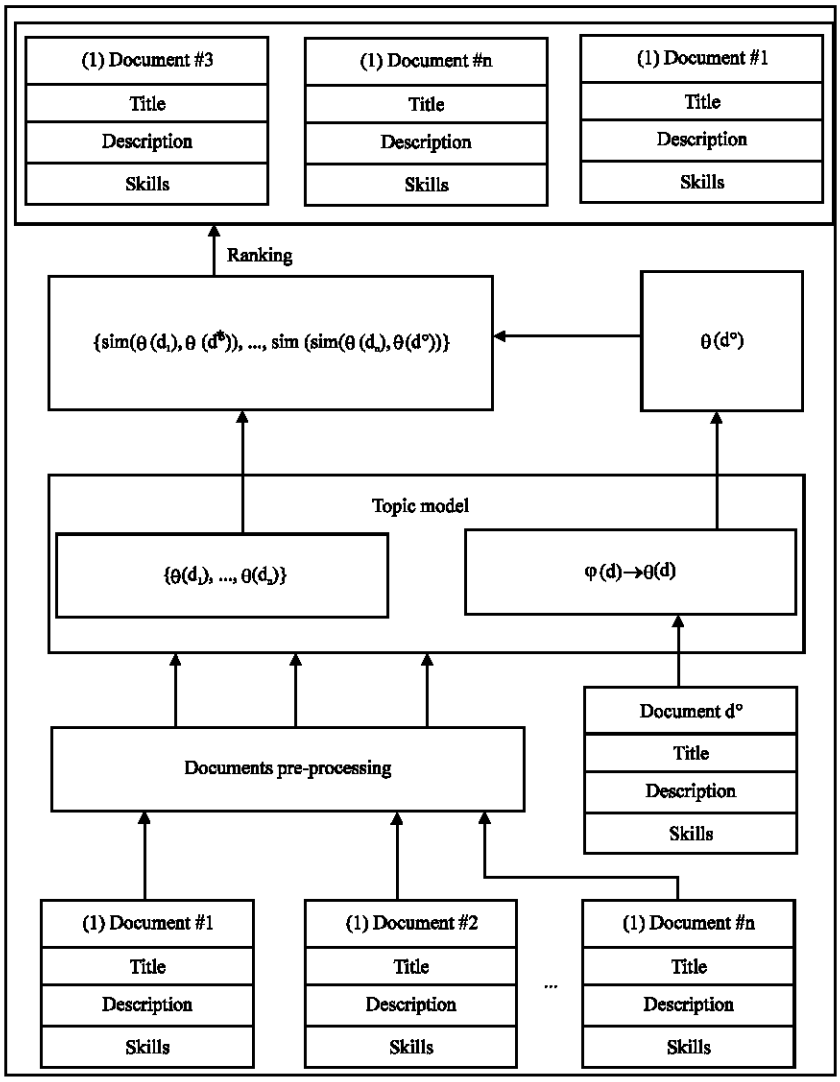

Fig. 2: Recommendation generation diagram proposed

recommendations for the search for IT service provider is reduced to finding a set of documents (d_1, ..., d_n) with topic distribution similar to that in documents (textual descriptions of project research or IT service provider) requiring a recommendation to be generated, i.e., documents with vectors $\theta(d_1)$, ..., $\theta(d_n)$ similar to target vector $\theta(d^*)$. They should be ranked by degree of similarity, more similar documents being supposed to more relevant for the user and should be shown at the beginning of the list of recommendations.

Many methods of determining the degree of vector similarity in a multidimensional space are known (Euclidian metric, manhattan distance, Mahalonobis distance). The general problem of estimating vector closeness is the 'curse of dimension': metric calculation has to be performed n times (n being the number of documents in the database except target document) for one target document in case of recommendation system. The problem is particularly topical when recommendation generation requests are intensively coming and the number of similarity measure calculations required is determined by the next formula:

$$C_n^2 = \frac{n!}{2(n-2)!}$$

where, n is the number of documents in the collection. One of the options of reducing the criticality of this problem is using an inverted index, i.e., essentially keeping a set of documents for each topic such that probability distribution is positive in them for this topic at

the stage of matrix θ generation. Recommendations are generated for document d* to assess similarity by selecting only documents contained in positions in the inverted index where probability distribution is positive in document d*. Since, matrix θ is heavily sparsed during regularization, the number of similarity measure calculations is significantly decreased (Ianina *et al*., 2017). Nevertheless, the number of calculations required can remain rather large (for large document collections). In this connection, the researcher suggests that only documents where topic probability is higher than average be selected to the inverted index for the topic. Average topic probability values are expedient to be calculated once in the process of learning and then the use of this indicator will not affect the computational complexity of the similarity calculation algorithm. The researcher suggests that documents be only selected for assessing similarity with the target document from topics such that their probability in the target document is higher than average probability value in the document. Since, vectors θ(d) are normalized by unit, average probability is a constant and is inversely proportional to the number of model topics, therefore, the use of this indicator does not affect the computational complexity of the similarity calculation algorithm. A formal presentation of the selection methods proposed is presented in the following formulas:

$$I(T_i) = \sum_{j=1}^{n} d_j(\theta_i(d_j) > \sum_{k=1}^{n} \theta_i(d_k)/n)$$

Where:
d_j = Collection document j
$\theta_i(d)$ = Probability value for topic i in document d

$$R(d^*) = \sum_{i=1}^{T-T_b} d(\theta_i > \sum_{k=1}^{T} \theta(d^*)/T) = \sum_{i=1}^{T-T_b} d(\theta_i > 1/T)$$

Where:
T_i = Number of subject topics and background topics
$\theta_i(d)$ = Number of background topics

The use of the selection methods proposed enables reducing the computational complexity of the similarity calculation algorithm for document vectors from database and target document due to reduction in the number of the calculation operations required. At the same time, exclusion of documents from the reference set does not significantly affect the quality of recommendations as only documents with comparatively low topic distribution probabilities are excluded which are irrelevant for the target document in the overwhelming majority of cases.

Despite availability of a great number of metrics for measuring vector closeness in a multidimensional space, the cosine measure of closeness essentially measuring the angle cosine is used most frequently (Sidorov *et al*., 2014). The cosine measure considers vector model attributes as independent and completely separate, although, some topics may be similar to each other in the model. In this connection, the researcher suggests that a 'soft' cosine measure taking into account similarity between topics be used to calculate the measure of vector closeness in the IT service provider search recommendation algorithm (Sidorov *et al*., 2014). The researcher suggests that the cosine measure by vectors of term distribution in topics be used to calculate similarity between topics. A formal presentation is described in the next formula:

$$\text{cosine}(T_1, T_2) = \frac{\sum_{i=1}^{n} \Phi_i(T_1)\Phi_i(T_2)}{\sqrt{\sum_{i=1}^{n} (\Phi_i(T_1))^2}\sqrt{\sum_{i=1}^{n} (\Phi_i(T_2))^2}}$$

Where:
n = Number of tokens
$\Phi_i(T)$ = Probability of Topic T for token I

The measure of similarity between topics is proposed to be calculated ones at the model learning stage. The use of this indicator will not affect the computational complexity of the document closeness calculation algorithm in that case.

If similarity between topics is used, the ultimate similarity calculation formula for two documents will take the form shown in the next formula:

$$\text{soft\_cosine}(d_1, d_2) = \frac{\sum_{i,j}^{N} s_{ij} d_{1i} d_{2j}}{\sqrt{\sum_{i,j}^{N} s_{ij} d_{1i} d_{1j}}\sqrt{\sum_{i,j}^{N} s_{ij} d_{2i} d_{2j}}}$$

Where:
S_{ij} = Similarity between topics i and j
d_{ni} = Probability of topic i for document d_n

Once, similarity measures have been calculated for all document selected, recommendations are shown to the user in the form of a list ranked in descending order for this measure.

The IT service provider search database is seeded with new documents over time. Model re-learning seems inexpedient on each appearance of a new document as it will increase the load on the computational system while the effect of an individual document on model structure is insignificant. In this connection, the researcher suggests that model re-learning be carried out (including the building of an inverted index and calculation of the measure of similarity between model topics) using one of the following methods: periodically (for example, weekly), on accumulation of a critical mass of documents that have not participated in model building (for example, 5% of participating documents) on reduction of external model quality indicators.

RESULTS AND DISCUSSION

Experiments have been carried out on a collection of textual descriptions in English including 12287 software development project research descriptions and 14344 of IT service provider descriptions.

A topic model containing 20 topics including three background ones was built. The regularization strategy was selected using the method of maximization of the weighted normalized sum of internal quality criteria. The strategy selected enabled improving internal model quality indicators as shown in Table 1.

The 10% of the total number of documents or 2664 documents were randomly selected as a test sample. Table 2 shows a comparison of the number of calculations required to calculate the measure of closeness between document vectors calculated for each document from the test sample (using the pre-selection proposed and without it).

Mean average precision at 10 (map@10) and Mean reciprocal rank at 10 (Chapelle *et al*., 2009) were used as ranking quality metrics in this study on the basis of binary expert assessments. Each document from the test sample was assigned a ranked list of recommendations consisting of 10 documents (list of IT service providers for project researchs). The experts were proposed to assess the relevance of documents from the ranked list to the target document using a binary scale (0 meaning non-relevant and 1 meaning relevant). As a result, each expert's assessment is a set of tuples of the form {(10, ..., 1) (00, ..., 1), ..., (10, ..., 0)}. The potency of the set equals the number of documents in the test sample and the length of each tuples equals the length of ranked list (10 in this case). Expert's assessment were averaged by each metric. The mean average precision at 10 proved to be equal to 0.73 and mean reciprocal rank at 10-1.74. It suggests that mean precision on the first 10 documents equals 73% and the user will most probably find the first relevant document in position 1.74 on the average (the positions themselves are obviously designated by natural numbers).

Table 1: Internal model quality indicators

Modality/Indicators	Without regularization	Optimal strategy	Δ (%)
Title			
Perplexity	222.88	280.15	26.13
Background score	0.06	0.06	0.00
Kernel contrast	0.96	0.99	3.13
Kernel purity	0.46	0.57	23.91
Is sparsity	0.88	0.96	9.09
Description			
Perplexity	1453.36	1748.12	20.28
Background score	0.31	0.31	0
Kernel contrast	0.97	0.98	1.03
Kernel purity	0.18	0.22	22.2
Is sparsity	0.83	0.93	12.05
Skills			
Perplexity	123.37	150	21.59
Background score	0.04	0.04	0
Kernel contrast	0.943	0.97	2.86
Kernel purity	0.4	0.49	22.5
Is sparsity	0.83	0.94	13.25
Is sparsity	0.03	0.84	2700

Table 2: The necessary number of document closeness measure calculations

Indicators	Without pre-selection	With pre-selection	Δ (%)
Average	11288	10239	-9
Mode	4495	3632	-19
Median	9653	7288	-24.5

The proposed approach to recommendation generation during IT service provider search for software development and maintenance project researches differs from known approaches in this sphere in that it requires no intellectual input processing. In contrast to known topic model-based recommendation algorithms, the approach is characterized by a lower computational complexity due to document pre-selection on the basis of topic distribution in them. In addition, the proposed approach takes into consideration topic similarity and distinction through the use of the cosine measure during document ranking in contrast to known topic modeling-based recommendation algorithms.

CONCLUSION

A ranking quality assessment performed using the proposed algorithm has shown good results which suggests that it can be used in industrial setting. The results obtained are supposed to be practically used to improve software development and maintenance tools during the use of IT outsourcing.

ACKNOWLEDGEMENT

The reported study was funded by RFBR according to the research project No. 18-01-00558.

REFERENCES

Al-Otaibi, S.T. and M. Ykhlef, 2012. A survey of job recommender systems. Intl. J. Phys. Sci., 7: 5127-5142.

Burke, R., 1999. Integrating knowledge-based and collaborative-filtering recommender systems. Proceedings AAAI-99 Workshop on AI and Electronic Commerce (AIEC99), July 18, 1999, AAAI, Orlando, Florida, pp: 69-72.

Chapelle, O., D. Metlzer, Y. Zhang and P. Grinspan, 2009. Expected reciprocal rank for graded relevance. Proceedings of the 18th ACM Conference on Information and Knowledge Management, November 2-6, 2009, Hong Kong, China, ISBN:978-1-60558-512-3, pp: 621-630.

Ghazanfar, M. and A. Prugel-Bennett, 2010. Building switching hybrid recommender system using machine learning classifiers and collaborative filtering. IAENG Int. J. Comput. Sci., 37: 272-287.

Gomzin A.G. and A.V. Korshunov, 2012. Recommendation systems: A survey of modern approaches. Proc. Inst. Syst. Program., 22: 401-418.

Ianina, A., L. Golitsyn and K. Vorontsov, 2017. Multi-objective topic modeling for exploratory search in tech news. Proceedings of the 2017 International Conference on Artificial Intelligence and Natural Language, September 20-23, 2017, Springer, Berlin, Germany, pp: 181-193.

Kochedykov, D., M. Apishev, L. Golitsyn and K. Vorontsov, 2017. Fast and modular regularized topic modelling. Proceedings of the 2017 21st Conference on Open Innovations Association (FRUCT), Novmber 6-10, 2017, IEEE, Helsinki, Finland, ISBN:978-1-5386-2136-3, pp: 182-193.

Pazzani, M.J. and D. Billsus, 2007. Content-Based Recommendation Systems. In: The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., A. Kobsa and W. Nejdl (Eds.). Springer, New York, pp: 325-341.

Salov, N., 2017. Feasibility of economical information system lifecycle tasks and performers co-clustering structure reconstruction based on their textual description. Trans. Bus. Russia, 4: 29-33.

Sidorov, G., A. Gelbukh, H. Gomez-Adorno and D. Pinto, 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. Computacion Sistemas, 18: 491-504.

Vorontsov K.V. and A.A. Potapenko, 2012. Regularization, robust and sparse of probability topic models. Comput. Stud. Model., 4: 693-706.

Vorontsov, K.V., 2014. Additive regularization for topic models of text collections. Dokl. Math., 89: 301-304.